On Testing Dependence between Time to Failure and Cause of Failure when Causes of Failure Are Missing

Isha Dewan¹, Sangita Kulathinal^{2*}

¹ Indian Statistical Institute, New Delhi, India, ² Indic Society for Education and Development, Nashik, India

The hypothesis of independence between the failure time and the cause of failure is studied by using the conditional probabilities of failure due to a specific cause given that there is no failure up to certain fixed time. In practice, there are situations when the failure times are available for all units but the causes of failures might be missing for some units. We propose tests based on U-statistics to test for independence of the failure time and the cause of failure in the competing risks model when all the causes of failure cannot be observed. The asymptotic distribution is normal in each case. Simulation studies look at power comparisons for the proposed tests for two families of distributions. The one-sided and the two-sided tests based on Kendall type statistic perform exceedingly well in detecting departures from independence.

Citation: Dewan I, Kulathinal S (2007) On Testing Dependence between Time to Failure and Cause of Failure when Causes of Failure Are Missing. PLoS ONE 2(12): e1255. doi:10.1371/journal.pone.0001255

INTRODUCTION

We consider a competing risks set-up where a unit is subject to two disjoint risks of failure and each unit ultimately fails due to either of the two risks. We do not allow simultaneous failures due to both the risks. The observations are made on the time to failure T and an identifier of the risk $\delta = j$ if the failure is due to the risk j , $j = 0, 1$. Let the joint distribution of (T, δ) be specified by the subsurvival functions $S_j(t) = P(T \geq t, \delta = j)$, or the subdistribution function given by $F_j(t) = P(T < t, \delta = j)$, $j = 0, 1$. The survival function and the distribution function of T are, respectively, given by

$$S(t) = P(T \geq t) = S_0(t) + S_1(t),$$

$$F(t) = P(T < t) = F_0(t) + F_1(t).$$

We assume that the subsurvival functions are continuous. Note that the distribution of the failure time and the cause of failure is specified using the observable variables (T, δ) .

Let the conditional probability of failure due to the first risk given that there is no failure up to time t be given as

$$\Phi_1(t) = P(\delta = 1 | T \geq t) = S_1(t)/S(t),$$

whenever $S(t) > 0$. These probabilities were introduced while studying failure and preventive maintenance in a censoring setting where the interest is in the distribution of the failure time which would have been observed in the absence of preventive maintenance [1]. Another conditional probability of interest is

$$\Phi_0^*(t) = P(\delta = 0 | T < t) = F_0(t)/F(t),$$

whenever $F(t) > 0$.

Under independence $S_j(t) = S(t)P(\delta = j)$ and hence, T and δ can be studied separately. Thus, the hypothesis of equality of subsurvival functions reduces to testing whether $P(\delta = 1) = P(\delta = 0) = 1/2$, a Bernoulli trial with success probability half. Hence a two-dimensional problem reduces to one-dimensional problem. The dependence between the failure time T and the cause of failure δ in terms of the above two conditional probability functions was studied in [2]. Below we give formal proofs of two results, which were stated in [2] on the independence and positive

quadrant dependent (PQD) structure of (T, δ) in terms of these conditional probabilities.

Lemma 1: T and δ are independent if and only if $\Phi_1(t) = \Phi_1(0) = \phi$, for all t where $\phi = P(\delta = 1)$.

Proof: When T and δ are independent their joint distribution is written as the product of the marginal distributions. Hence,

$$S_1(t) = P(\delta = 1)P(T \geq t) = \phi S(t),$$

$$\Phi_1(t) = \frac{S_1(t)}{S(t)} = \frac{\phi S(t)}{S(t)} = \phi,$$

where $\phi = P(\delta = 1)$. Similarly when $\Phi_1(t)$ does not depend on t then $\Phi_1(t) = \Phi_1(0)$ and $\Phi_1(0) = S_1(0)/S(0) = P(\delta = 1) = \phi$.

This in turn implies that $S_1(t) = \phi S(t)$ which is the product of the marginal distributions of δ and t . Also, $S_0(t) = S(t) - S_1(t) = (1 - \phi)S(t)$. Hence the result.

The independence of T and δ is also equivalent to $\Phi_0^*(t) = \Phi_0^*(0)$, for all t . A simple and easily checked dependence structure is positive quadrant dependent (PQD) indicating positive association between two random variables.

Definition 1: Random variables X and Y are Positive Quadrant Dependent (PQD) if the following inequality holds:

$$P(X \leq x, Y \leq y) \geq P(X \leq x)P(Y \leq y), \text{ for all } x \text{ and } y.$$

In our case, because δ takes only two values 0 and 1, T and δ are

Academic Editor: Enrico Scalas, University of East Piedmont, Italy

Received July 11, 2007; Accepted August 3, 2007; Published December 5, 2007

Copyright: © 2007 Dewan, Kulathinal. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* To whom correspondence should be addressed. E-mail: sangita.kulathinal@inseed.org

PQD if the following inequality holds:

$$P(T \leq t, \delta = 0) \geq P(T \leq t)P(\delta = 0), \text{ for all } t.$$

This is because $P(T \leq t, \delta \leq 0) = P(T \leq t, \delta = 0)$, $P(T \leq t, \delta \leq 1) = P(T \leq t)$, and $P(\delta \leq 1) = 1$. Hence the required inequality always holds for $\delta = 1$.

Lemma 2: T and δ are PQD if and only if $\Phi_1(t) \geq \Phi_1(0) = \phi$, for all t .

Proof: When T and δ are PQD then

$$P(T \leq t, \delta = 0) \geq P(T \leq t)P(\delta = 0) \text{ for all } t.$$

Note that

$$\begin{aligned} P(T \leq t, \delta = 0) &= P(\delta = 0) - P(T > t, \delta = 0) \\ &= 1 - P(\delta = 1) - (P(T > t) - P(T > t, \delta = 1)) \\ &= 1 - \phi - (S(t) - S_1(t)), \end{aligned}$$

and $P(T \leq t)P(\delta = 0) = (1 - S(t))(1 - \phi)$. Substituting these identities in the above inequality, we get

$$1 - \phi - (S(t) - S_1(t)) \geq (1 - S(t))(1 - \phi)$$

$$S_1(t) \geq \phi S(t) \text{ or } \Phi_1(t) = \frac{S_1(t)}{S(t)} \geq \phi.$$

Hence the result.

Note that T and δ are *positive quadrant dependent* (PQD) is also equivalent to $\Phi_0^*(t) \geq \Phi_0^*(0)$, for all $t > 0$. Various hypothesis testing problems of checking independence of T and δ against various alternatives specifying dependence structures are considered and U-statistics are derived when the complete data on all n units are available [2]. However, in many practical situations the experimenter may have information on failure times for all the individuals but on the causes of failures only for some.

In mortality follow-up study, the causes of death are obtained from the death certificates. The problem of causes of death missing on death certificates is well-known. This may occur due to various reasons like doctor's strike, autopsy not performed and hence, no knowledge of Definite underlying cause of death, and no legal requirement of mentioning an underlying cause of death on the death certificate. The present work is motivated by a follow-up study on mortality where the underlying causes of death were missing for nearly 20% patients but the times of death were known for all. Similar situation arises in engineering fields where series systems are tested for failure due to various components, possibly under accelerated life testing. In this case, a thorough autopsy of failed system is required to identify the failed component(s) which leads to the system failure. Such information may not be available for all failed systems due to financial or logistic reasons. In motorcycle fatalities study, 40% of the death certificates had either partial or no information on underlying causes of death [3].

An example from animal bioanalysis where all causes were not available was considered [4]. Likelihood based estimation in case of missing causes of failure have been studied [5,6,7,8,9]. A modified log rank test for competing risks with missing failure type was also considered [10]. The maximum likelihood estimators and minimum variance unbiased estimators of the parameters of exponential distribution for the missing case were obtained [11] and their approximate and asymptotic properties were discussed and confidence intervals were derived [12].

In this paper, we consider the problem of testing

$$H_0 : T \text{ and } \delta \text{ are independent}$$

against various alternative hypotheses characterising the dependence structure of T and δ , which are:

$$H_1 : T \text{ and } \delta \text{ are not independent}$$

$$H_2 : T \text{ and } \delta \text{ are positive quadrant dependent}$$

when causes are missing for some units. Let (T_i, δ_i) , $i = 1, \dots, n$ be the competing-risks data available on n individuals. Here, we consider a situation when δ_i may not be observed always i.e., it may be missing for some units. Let O_i be an indicator variable which takes value one if δ_i is observed and zero if δ_i is missing. Let $p = pr(O_i = 1)$ be the probability that δ is observed. We assume that δ_i are missing completely at random and hence O_i is independent of (T_i, δ_i) . The fact that the cause of failure will be observed or not will have no bearing on the actual cause. Similar assumptions are made in [9,13].

We extend some of the tests based on U-statistics proposed in [2] to the case when δ 's are not observed for all the units. We carry out simulation studies for comparing the power of the tests for two families of distributions. We also apply the proposed tests to the data on failure of switches given in [14] by artificially creating missing causes of failure. The proposed tests perform satisfactorily and the use of the data on the failure times even when corresponding causes are missing is recommended.

RESULTS

We apply the proposed tests U_{km} , U_{PQDm} and U_{km1} , the one-sided test based on U_{km} to simulated data from two parametric families of distributions and evaluate empirical powers. We also apply the tests to a real data. The computations were done using SAS [15] and the source codes and a brief guide on how to use the SAS codes are provided in the supplementary material (Text S1, Text S2, Text S3 and Text S4).

Example 1: Parametric family of distributions [2]

Consider the parametric family of distributions proposed in [2]

$$F_1(t) = \phi F^a(t),$$

$$F_0(t) = F(t) - \phi F^a(t),$$

where $1 \leq a \leq 2$, $0 \leq \phi \leq 0.5$ and $F(t)$ is a proper distribution function. Note that $P(\delta = 1) = \phi$ and

$$\Phi_1(t) = \frac{\phi(1 - F^a(t))}{1 - F(t)}$$

which is an increasing function of t . For $a = 1$, $\Phi_1(t) = \phi$, that is, T and δ are independent and for $1 < a \leq 2$, $\Phi_1(t) > \phi$, that is T and δ are PQD, and hence H_2 holds. Let $F(t) = 1 - \exp\{-\lambda t\}$ be the overall distribution function. We simulated random samples by varying n , p , and (λ, a, ϕ) from the above distribution. Empirical level of significance and power were calculated by using 1000 replications for each combination of n , and (λ, a, ϕ) . Table 1 gives the empirical level of significance and Table 2 gives the empirical power of the three tests based on U_{km} , U_{km1} and U_{PQDm} .

From the two tables it is clear that modified test statistic U_{km} attains its level when roughly half of the failures are likely to be due

Table 1. Empirical level of significance of the three U-tests for a family of distributions of Dewan *et al.* (2004) $(\lambda, a) = (1, 1)$

ϕ	p	$n = 25$			$n = 50$			$n = 100$		
		U_{km}	U_{km1}	U_{PQDm}	U_{km}	U_{km1}	U_{PQDm}	U_{km}	U_{km1}	U_{PQDm}
0.5	1	0.057	0.054	0	0.061	0.049	0.001	0.052	0.047	0.002
0.5	0.9	0.055	0.048	0.002	0.062	0.052	0.001	0.047	0.053	0.004
0.5	0.8	0.054	0.047	0.007	0.049	0.061	0.008	0.05	0.049	0.004
0.2	1	0.069	0.061	0.001	0.052	0.049	0.001	0.051	0.06	0.01
0.2	0.9	0.078	0.059	0.001	0.061	0.052	0.002	0.052	0.059	0.005
0.2	0.8	0.092	0.06	0.005	0.057	0.054	0.004	0.047	0.064	0.002

doi:10.1371/journal.pone.0001255.t001

to the first cause. The conclusions are valid even when 20% of the failure causes are not available. The power increases with increase in values of a and also with increase in sample size. The test has very good power even when $a = 1.5$. One should keep in mind the fact that the alternative of no independence is extremely general.

However, the test based on one-sided version of the Kendall's τ , U_{km1} performs much better than the test based on U_{PQDm} for testing H_0 against H_2 . It was observed in [2] that the test U_{PQDm} , when $p = 1$ is extremely conservative and also inefficient. The entries for this test in the two tables confirm this observation. But given the fact that the level of significance attained is very low, it is able to detect alternatives reasonably well.

Example 2: Random sign censoring model [1]

A random sign censoring (RSC), also known as an age-dependent censoring, is a model in which the lifetime of a unit (X) is censored by $Z = X - W\eta$, where W , $0 < W < X$ is the time at which a warning is emitted by the unit before its failure, and η is a random variable taking values $\{-1, 1\}$ and is independent of X . Hence $\eta = 1$ would lead to the censoring of the life time at $X - W$ giving $T = Z$ and $\delta = 0$ and $\eta = -1$ will lead to the observation of complete lifetime X , giving $T = X$ and $\delta = 1$. Assume that X has exponential distribution with parameter λ . In this case, $P(Z \geq t, Z < X) = P(X - W \geq t, \eta = 1)$ and $P(X \geq t, X < Z) = P(X \geq t, \eta = -1)$. This gives $\Phi_1(t) = P(T \geq t, \delta = 1) / P(T \geq t) = P(X \geq t, \eta = -1) / P(X \geq t, X - W\eta \geq t)$.

Here $P(\eta = -1) = 1 - P(\eta = 1) = P(\delta = 1) = \phi$. When X is exponentially distributed with parameter λ and $W = aX$, $0 < a < 1$,

$$\Phi_1(t) = (1 + (1 - \phi) / \phi \exp\{-\lambda t(a / (1 - a))\})^{-1},$$

leading to the increasing (and hence PQD) nature of $\Phi_1(t)$ in t . As a goes to zero $\Phi_1(t)$ goes to ϕ , that is T, δ are independent. Hence, to evaluate the empirical level of significance we choose a very close to zero.

The value a close to zero corresponds to independence of T and δ and $a > 0$ gives $\Phi_1(t)$ as an increasing function of t implying T and δ are PQD. For simulation purposes, we consider two values of $a = 0.00001$, and 0.5. Test based on U_{km} almost attains its level even for sample sizes as small as $n = 25$ as can be seen from Table 3. This test has good power for $n = 100$. The test based on U_{km1} has a slightly higher level as well as higher power. But the test based on U_{PQDm} is a very conservative test. It has low empirical power even for $n = 100$. One-sided test based on U_{km1} is definitely a better choice for detecting PQD alternatives.

Example 3: Nair's data revisited [14]

Here we consider the data on the failure of 37 switches due to one of the two possible causes of failures published in [14]. These data were analysed in [2] and it was shown that the failure time (T) and the cause of failure (δ) of switches were not independent. Also, the conditional probability of failure due to cause A, $\Phi_1(t)$ was shown to be larger than ϕ and hence T and δ were PQD.

We calculate three test statistics for the entire data on 37 switches as earlier. We also artificially create missing data on the cause of failure for varying values of p and repeating it for 1000 times to evaluate the empirical powers of the test statistics.

The hypothesis of independence of T and δ , H_0 is rejected against H_1 at $\alpha = 5\%$ level of significance using U_{km} (the value of the test statistic is 2.70 which is larger than 1.96) and the one-sided test, U_{km1} (the value of the test statistic is 2.70 which is larger than 1.64) rejects the hypothesis H_0 against H_2 at $\alpha = 5\%$ level of significance. However, the test based on PQD, U_{PQDm} (the value of the test statistic is 1.35 which is smaller than 1.64) does not reject the hypothesis H_0 against H_2 at $\alpha = 5\%$ level of significance. Table 4 shows the empirical powers of the tests for various values of p .

As seen earlier with the simulated data, the test U_{km1} performs well even when 60% of the causes are missing. The power of U_{PQDm} test is unsatisfactory.

Table 2. Empirical power of the three U-tests for a family of distributions of Dewan *et al.* (2004) $\lambda = 1$ and $\phi = 0.5$

ϕ	p	$n = 25$			$n = 50$			$n = 100$		
		U_{km}	U_{km1}	U_{PQDm}	U_{km}	U_{km1}	U_{PQDm}	U_{km}	U_{km1}	U_{PQDm}
1.5	1	0.421	0.546	0.049	0.690	0.812	0.201	0.942	0.974	0.575
1.5	0.9	0.394	0.505	0.066	0.652	0.767	0.201	0.916	0.955	0.504
1.5	0.8	0.365	0.489	0.079	0.609	0.711	0.208	0.894	0.938	0.453
1.8	1	0.765	0.863	0.183	0.967	0.988	0.646	0.999	0.999	0.974
1.8	0.9	0.713	0.802	0.181	0.947	0.975	0.545	0.998	0.999	0.917
1.8	0.8	0.653	0.768	0.191	0.921	0.965	0.485	0.997	0.999	0.836

doi:10.1371/journal.pone.0001255.t002

Table 3. Empirical power of the three U-tests for random sign censoring model ($\lambda = 1$)

(a, ρ)	ρ	$n = 25$			$n = 50$			$n = 100$		
		U_{km}	U_{km1}	U_{PQDm}	U_{km}	U_{km1}	U_{PQDm}	U_{km}	U_{km1}	U_{PQDm}
$(10^{-5}, 0.5)$	1	0.057	0.067	0.002	0.061	0.063	0	0.052	0.065	0
$(10^{-5}, 0.5)$	0.9	0.055	0.065	0.001	0.062	0.059	0.001	0.047	0.053	0
$(10^{-5}, 0.5)$	0.8	0.054	0.059	0.005	0.049	0.059	0.008	0.05	0.048	0.006
$(10^{-5}, 0.7)$	1	0.056	0.057	0	0.067	0.063	0	0.056	0.057	0
$(10^{-5}, 0.7)$	0.9	0.063	0.069	0.005	0.059	0.064	0.003	0.055	0.06	0.003
$(10^{-5}, 0.7)$	0.8	0.062	0.067	0.009	0.057	0.063	0.006	0.052	0.056	0.000
$(0.5, 0.5)$	1	0.352	0.474	0.029	0.577	0.702	0.11	0.85	0.923	0.381
$(0.5, 0.5)$	0.9	0.321	0.442	0.05	0.513	0.647	0.135	0.793	0.884	0.358
$(0.5, 0.5)$	0.8	0.286	0.41	0.066	0.473	0.609	0.121	0.751	0.848	0.306
$(0.5, 0.7)$	1	0.32	0.423	0.029	0.489	0.61	0.091	0.759	0.847	0.303
$(0.5, 0.7)$	0.9	0.266	0.378	0.045	0.449	0.564	0.105	0.705	0.811	0.249
$(0.5, 0.7)$	0.8	0.229	0.348	0.056	0.396	0.516	0.112	0.655	0.755	0.233

doi:10.1371/journal.pone.0001255.t003

DISCUSSION

Testing independence between the failure time T and the cause of failure δ is often important because of reduction in dimensionality and possibility of studying T and δ separately. The available tests use only completely observed data on T and δ . One cannot avoid missing data situation in practice and hence, the issue of the effect missing observations on the existing tests needs to be addressed.

From the simulation studies it is clear that the two-sided test, U_{km} is performing well for both the families of distributions for sample sizes as small as 25 and when 20% of the causes of failure are not known. These observations can be made from Table 1, Table 2 and Table 3 with attained level of significance and high empirical power. The empirical powers for all the three tests are higher in the case of the parametric family of distributions of Example 1 compared to RSC model of Example 2 for all sample sizes. The performance of the one-sided test, U_{km1} based on Kendall's τ is clearly superior to the U_{PQDm} as demonstrated by Table 1, Table 2 and Table 3. Even when all causes are known it observed that the test based on Kendall's τ is four times more efficient than the test based on U_{PQD} [2]. Even in the case of missing causes, we recommend the use of U_{km1} for testing independence against PQD. One obvious reason is that Kendall's τ uses information on (T, δ) for each pair of observations. Similar observations are made on the basis of real data analysis of Example 3 (Table 4).

The failure times with missing information on causes of failures also provide useful information regarding departures from independence of T and δ , and hence, omitting such observations from the analysis may result in loss of efficiency. For this reason,

Table 4. Empirical power of the three U-tests for Nair's data (1993)

ρ	U_{km}	U_{km1}	U_{PQDm}
0.9	0.962	0.993	0.127
0.8	0.841	0.945	0.167
0.7	0.705	0.865	0.18
0.6	0.578	0.758	0.187
0.5	0.454	0.638	0.172

doi:10.1371/journal.pone.0001255.t004

the analysis may not be based on only the complete data on both time and causes of failures (with reduced sample size, which is random). This article is the first attempt of its kind to carry out the tests for independence under the assumption of missing completely at random. How the tests perform under the assumption of missing at random or even informative missingness remains an open research problem.

MATERIALS AND METHODS

General dependence between T and δ

First we consider the problem of testing $H_0 : \Phi_1(t) = \phi$, for all t against $H_1 : \Phi_1(t)$ is not a constant, where $\Phi_1(t) = P(\delta = 1 | T \geq t) = S_1(t)/S(t)$, and $\phi = \Phi_1(0) = P(\delta = 1)$. Recall that a pair (T_i, δ_i) and (T_j, δ_j) is a concordant pair if $T_i > T_j$, $\delta_i = 1$, $\delta_j = 0$ or $T_i < T_j$, $\delta_i = 0$, $\delta_j = 1$ and is a discordant pair if $T_i > T_j$, $\delta_i = 0$, $\delta_j = 1$ or $T_i < T_j$, $\delta_i = 1$, $\delta_j = 0$. The U -statistic based on the idea of concordance and discordance pairs or Kendall's τ is

$$U_k = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \psi_k(T_i, \delta_i, T_j, \delta_j),$$

where

$$\psi_k(T_i, \delta_i, T_j, \delta_j) = \begin{cases} 1 & \text{if } T_i > T_j, \delta_i = 1, \delta_j = 0 \\ & \text{or } T_i < T_j, \delta_i = 0, \delta_j = 1 \\ -1 & \text{if } T_i > T_j, \delta_i = 0, \delta_j = 1 \\ & \text{or } T_i < T_j, \delta_i = 1, \delta_j = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where the subscript k indicates that the U -statistic is defined using the idea of Kendall's τ [2]. If δ is missing for some units then $\psi_k(T_i, \delta_i, T_j, \delta_j)$ cannot be defined for all pairs. In Table 5 m indicates that δ is not observable and ? indicates the cases when ψ_k is not defined.

Note that when $T_i > T_j$ and $\delta_i = 1$, but δ_j is missing, $\psi_k(T_i, \delta_i, T_j, \delta_j)$ will take value 1 if $\delta_j = 0$ and value 0 if $\delta_j = 1$. Hence, in order to retrieve the best possible information we assign weight $(1+0)/2 = 1/2$ in this case. Similarly, when $T_i > T_j$ and $\delta_i = 0$, but δ_j is missing, $\psi_k(T_i, \delta_i, T_j, \delta_j)$ will take value -1 if $\delta_j = 1$ and value 0 if $\delta_j = 0$. Hence, we assign value $-1/2$ to the kernel in this case.

Table 5. Values taken by the kernel ψ_k for various combinations of the pairs (T_i, δ_i) and (T_j, δ_j)

(δ_i, δ_j)	(1,1)	(1,0)	(1,m)	(0,1)	(0,0)	(0,m)	(m,1)	(m,0)	(m,m)
$T_i > T_j$	0	1	?	-1	0	?	?	?	?
$T_i \leq T_j$	0	-1	?	1	0	?	?	?	?

doi:10.1371/journal.pone.0001255.t005

Now, we redefine the kernel when some observations on δ are missing as follows

$$\Psi_{km}(T_i, \delta_i, O_i, T_j, \delta_j, O_j) = \begin{cases} 1 & \text{if } T_i > T_j, \delta_i = 1, \delta_j = 0, O_i = 1, O_j = 1 \\ & \text{or } T_i < T_j, \delta_i = 0, \delta_j = 1, O_i = 1, O_j = 1 \\ 1/2 & \text{if } T_i > T_j, \delta_i = 1, O_i = 1, O_j = 1 \\ & \text{or } T_i < T_j, \delta_i = 0, O_i = 1, O_j = 0 \\ & \text{if } T_i > T_j, \delta_j = 0, O_i = 0, O_j = 1 \\ & \text{or } T_i < T_j, \delta_j = 1, O_i = 0, O_j = 1 \\ -1/2 & \text{if } T_i > T_j, \delta_i = 0, O_i = 1, O_j = 0 \\ & \text{or } T_i < T_j, \delta_i = 1, O_i = 1, O_j = 0 \\ & \text{if } T_i > T_j, \delta_j = 1, O_i = 0, O_j = 1 \\ & \text{or } T_i < T_j, \delta_j = 0, O_i = 0, O_j = 1 \\ -1 & \text{if } T_i > T_j, \delta_i = 0, \delta_j = 1, O_i = 1, O_j = 1 \\ & \text{or } T_i < T_j, \delta_i = 1, \delta_j = 0, O_i = 1, O_j = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Here the subscript m indicates missing data situation. Define U_{km} as the corresponding U -statistic

$$U_{km} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \psi_{km}(T_i, \delta_i, O_i, T_j, \delta_j, O_j).$$

Then the expectations of U_{km} is given by

$$\begin{aligned} E(U_{km}) &= E(\psi_{km}(T_i, \delta_i, O_i, T_j, \delta_j, O_j)) \\ &= 2p^2\phi + 4p^2 \int_0^\infty S(t) dS_1(t), \\ &= 0, \text{ under } H_0, \end{aligned}$$

and the asymptotic variance of $\sqrt{n}(U_{km} - E(U_{km}))$ under H_0 , denoted as $Var(U_{km})$ is

$$Var(U_{km}) = \sigma_{km}^2 = \frac{4}{3}p^2\phi(1-\phi) + \frac{1}{3}p(1-p).$$

Note that when $p=1$, the variance simplifies to $(4/3)\phi(1-\phi)$, which is given in [2]. Also, $E(U_{km}) \neq 0$ under H_1 . From the central limit theorem of U -statistics [16] (see Text S5), it follows that U_{km} has asymptotic normal distribution for large n .

Theorem 1: Under H_0 , $\sqrt{n}U_{km}$ converges in distribution to $N(0, \sigma_{km}^2)$ as $n \rightarrow \infty$, where $\sigma_{km}^2 = (4/3)p^2\phi(1-\phi) + (1/3)p(1-p)$.

We refer to the supplementary material (Text S6) for the explicit derivation of $E(U_{km})$, $Var(U_{km})$ and the proof of Theorem 1.

In practice, p and ϕ are generally unknown and can be replaced by their consistent estimators, $\hat{p} = \sum_{i=1}^n I(O_i = 1)/n$, and $\hat{\phi} = \sum_{i=1}^n I(\delta_i = 1)/n$, respectively. A test procedure for testing H_0 against H_1 is then: reject H_0 at 100 $\alpha\%$ level of significance if $|\sqrt{n}U_{km}/\hat{\sigma}_{km}|$ is larger than $z_{1-\alpha}$, cut-off point of standard normal distribution, where $\hat{\sigma}_{km}^2$ is a consistent estimator of σ_{km}^2 got by replacing p and ϕ by $\hat{p}$ and $\hat{\phi}$.

For computational purposes, it is necessary to express U_{km} as a function of ranks. Let $n_1 = \sum_{i=1}^n I(\delta_i = 1, O_i = 1)$, $n_2 = \sum_{i=1}^n I(\delta_i = 0, O_i = 1)$, and $n_3 = \sum_{i=1}^n I(O_i = 0)$ represent numbers of observations in three groups-causes are observed to be 1, causes are observed to be 0 and causes are not observed, respectively. Let the corresponding ordered times in each group be represented by $X_{(1)}, X_{(2)}, \dots, X_{(n_1)}$, $Y_{(1)}, Y_{(2)}, \dots, Y_{(n_2)}$, and $Z_{(1)}, Z_{(2)}, \dots, Z_{(n_3)}$, respectively. Let R_i denote the combined rank of $X_{(i)}$ in the ordered arrangement of (n_1+n_2) samples of type X and Y , S_i denote the combined rank of $X_{(i)}$ in the ordered arrangement of (n_1+n_3) samples of type X and Z , and Q_i denote the combined rank of $Y_{(i)}$ in the ordered arrangement of (n_2+n_3) samples of type Y and Z . Then, the number of observations for which $\psi_{km}(\cdot)$ takes value

$$\#(+1) = \sum_{i=1}^{n_1} (R_i - i) O_i \delta_i,$$

$$\#(-1) = n_1 n_2 - \sum_{i=1}^{n_1} (R_i - i) O_i \delta_i,$$

$$\#(+1/2) = \sum_{i=1}^{n_1} (S_i - i) O_i \delta_i + n_2 n_3 - \sum_{i=1}^{n_2} (Q_i - i) O_i (1 - \delta_i),$$

$$\#(-1/2) = n_1 n_3 - \sum_{i=1}^{n_1} (S_i - i) O_i \delta_i + \sum_{i=1}^{n_2} (Q_i - i) O_i (1 - \delta_i).$$

Thus, the expression U_{km} in terms of ranks is

$$\begin{aligned} \binom{n}{2} U_{km} &= 2 \sum_{i=1}^{n_1} (R_i - i) - n_1 n_2 + \frac{1}{2} \left[2 \sum_{i=1}^{n_1} (S_i - i) - \right. \\ &\quad \left. 2 \sum_{i=1}^{n_2} (Q_i - i) + n_2 n_3 - n_1 n_3 \right] = 2 \sum_{i=1}^{n_1} (R_i - i) + \\ &\quad \sum_{i=1}^{n_1} (S_i - i) - \sum_{i=1}^{n_2} (Q_i - i) - n_1 n_2 + \frac{n_2 n_3}{2} - \frac{n_1 n_3}{2}. \end{aligned} \quad (1)$$

Consider testing H_0 against H_2 . The U -statistic for testing H_0 against H_2 is

$$U_{PQD} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \psi_{PQD}(T_i, \delta_i, T_j, \delta_j),$$

Table 6. Values taken by the kernel ψ_{PQD} for various combinations of the pairs (T_i, δ_i) and (T_j, δ_j)

(δ_i, δ_j)	(1,1)	(1,0)	(1,m)	(0,1)	(0,0)	(0,m)	(m,1)	(m,0)	(m,m)
$T_i \leq T_j$	1	1	1	0	0	0	?	?	?
$T_i > T_j$	1	0	?	1	0	?	1	0	?

doi:10.1371/journal.pone.0001255.t006

where

$$\Psi_{PQD}(T_i, \delta_i, T_j, \delta_j) = \begin{cases} 1 & \text{if } T_i > T_j, \delta_i = 1 \\ & \text{or } T_j > T_i, \delta_j = 1 \\ 0 & \text{otherwise.} \end{cases}$$

This test was proposed in [2]. If δ is missing for some units then $\Psi_{PQD}(T_i, \delta_i, T_j, \delta_j)$ cannot be defined for all pairs. Table 6 shows the pairs for which the kernel is defined completely and also the cases where it is not defined.

As in the earlier subsection, we define a modified kernel to take into account missing causes as follows.

$$\Psi_{PQDm}(T_i, \delta_i, O_i, T_j, \delta_j, O_j) = \begin{cases} 1 & \text{if } T_i > T_j, \delta_i = 1, O_i = 1 \\ & \text{or } T_j > T_i, \delta_j = 1, O_j = 1 \\ 1/2 & \text{if } T_i > T_j, O_i = 0 \\ & \text{or } T_j > T_i, O_j = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let the corresponding U -statistic be U_{PQDm}

$$U_{PQDm} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \Psi_{PQDm}(T_i, \delta_i, O_i, T_j, \delta_j, O_j),$$

where the subscript m indicates missing data situation. The expectations of U_{PQDm} are given by

$$E(U_{PQDm}) = 2p \int_0^{\infty} S_1(t) dF(t) + \frac{(1-p)}{2} \\ = p\phi + \frac{(1-p)}{2}, \text{ under } H_0,$$

and $E(U_{PQDm}) > p\phi + \frac{(1-p)}{2}$ under H_2 . The asymptotic variance of $\sqrt{n}(U_{PQDm} - E(U_{PQDm}))$ under H_0 , denoted as $Var(U_{PQDm})$ is

$$Var(U_{PQDm}) = \sigma_{PQDm}^2 = \frac{4}{3}p^2\phi(1-\phi) + \frac{1}{3}p(1-p).$$

Note that when $p=1$, the variance simplifies to $(4/3)\phi(1-\phi)$, which is given in [2]. From the central limit theorem of U -statistics [16] (see Text S5), it follows that U_{PQDm} has asymptotic normal distribution for large n .

Theorem 2: Under H_0 , $\sqrt{n}(U_{PQDm} - E(U_{PQDm}))$ converges in distribution to $N(0, \sigma_{PQDm}^2)$, where $\sigma_{PQDm}^2 = (4/3)p^2\phi(1-\phi) + (1/3)p(1-p)$, as $n \rightarrow \infty$.

We refer to the supplementary material (Text S7) for the explicit derivation of $E(U_{PQDm})$, $Var(U_{PQDm})$ and the proof of Theorem 2.

We reject the null hypothesis for large values of $\sqrt{n}\hat{\sigma}_{PQDm}^{-1}(U_{PQDm} - E(\hat{U}_{PQDm}))$, where $E(\hat{U}_{PQDm})$ and $\hat{\sigma}_{PQDm}$ are obtained by replacing ϕ and p by their empirical estimators. As mentioned earlier, p and ϕ can be replaced by their consistent

estimators $\hat{p} = \sum_{i=1}^n I(O_i=1)/n$ and $\hat{\phi} = \sum_{i=1}^n I(\delta_i=1)/n$. Let R_i^*

denote the rank of T_i in the ordered observations $(T_1, T_2, \dots, T_n)$. Then, it is easy to see that

$$\binom{n}{2} U_{PQDm} = \sum_{i=1}^n (R_i^* - 1) O_i \delta_i + \frac{1}{2} \sum_{i=1}^n (R_i^* - 1) (1 - O_i).$$

Note that a one-sided test based on U_{km} , where H_0 is rejected for large values of U_{km} can also be used for testing H_0 against H_2 since $E(U_{km}) \geq 0$ under H_2 . In fact, the one-sided test uses data on both the T and δ in each pairwise comparison while U_{PQDm} uses only information on (T, δ) from one and T from the other in a pairwise comparison. We refer the one-sided test based on U_{km} as U_{km1} .

SUPPORTING INFORMATION

Text S1 SAS source code for Example 1

Found at: doi:10.1371/journal.pone.0001255.s001 (0.06 MB DOC)

Text S2 SAS source code for Example 2

Found at: doi:10.1371/journal.pone.0001255.s002 (0.06 MB DOC)

Text S3 SAS source code for Example 3

Found at: doi:10.1371/journal.pone.0001255.s003 (0.03 MB DOC)

Text S4 A short guide on the use of SAS codes

Found at: doi:10.1371/journal.pone.0001255.s004 (0.02 MB DOC)

Text S5 Central limit theorem for U -statistics

Found at: doi:10.1371/journal.pone.0001255.s005 (0.07 MB DOC)

Text S6 Derivation of $E(U_{km})$, $Var(U_{km})$ and proof of Theorem 1

Found at: doi:10.1371/journal.pone.0001255.s006 (0.14 MB DOC)

Text S7 Derivation of $E(U_{PQDm})$, $Var(U_{PQDm})$ and proof of Theorem 2

Found at: doi:10.1371/journal.pone.0001255.s007 (0.07 MB DOC)

ACKNOWLEDGMENTS

The authors would like to thank the Academic Editor for careful reading and comments which helped in improving the manuscript.

Author Contributions

Other: Contributed equally to the development of theory and simulation studies: SK ID. Contributed equally to the writing of the manuscript: SK ID.

REFERENCES

1. Cooke RM (1996) The design of reliability databases, part II. Reliability engineering and system safety 51: 209–223.
2. Dewan I, Deshpande JV, Kulathinal SB (2004) On testing dependence between time to failure and cause of failure via conditional probabilities. Scandinavian Journal of Statistics 31: 79–91.
3. Lapidus G, Braddock M, Schwartz R, Banco L, Jacobs L (1994) Accuracy of fatal motorcycle injury reporting on death certificates. Accident Anal. Prevention 26: 535–542.
4. Kodel RK, Chen JJ (1987) Handling cause of death in equivocal cases using the EM algorithm. Comm. Stats.-Theory and Methods 16: 2565–2603.
5. Dinse GE (1986) Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. J. Amer. Statist. Assoc. 81: 328–336.
6. Dewanji A (1992) A note on a test for competing risks with general missing pattern in failure types. Biometrics 79: 855–857.
7. Goetghebeur E, Ryan L (1995) Analysis of competing risks survival data when some failure types are missing. Biometrika 82: 821–833.
8. Dewanji A, Sengupta D (2003) Estimation of competing risks with general missing pattern in failure types. Biometrics 59: 1063–1070.
9. Lu K, Tsiatis A (2005) Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. Lifetime Data Anal. 11: 29–40.
10. Goetghebeur E, Ryan L (1990) A modified log rank test for competing risks with missing failure type. Biometrika 77: 207–211.
11. Miyakawa M (1984) Analysis of incomplete data in competing risks model. IEEE Transactions in Reliability 33: 293–296.
12. Kundu D, Basu S (2000) Analysis of incomplete data in presence of competing risks. J. Statist. Plann. Inference 87: 221–239.
13. Rubin DR (1976) Inference and missing data. Biometrika 63: 581–592.
14. Nair VN (1993) Bounds for reliability estimation under dependent censoring. International Stat. Review 61: 169–182.
15. SAS Institute Inc. SAS/STAT user's guide, version 8, 2000.
16. Puri ML, Sen PK (1971) Nonparametric methods in multivariate analysis: John Wiley, New York-London-Sydney.